

# **Finished Prokaryotic Genome Assemblies From a Low-cost Combination of Short and Long Reads**

- An ALLPATHS-LG recipe

Shuangye Yin

FSAF, June 2012

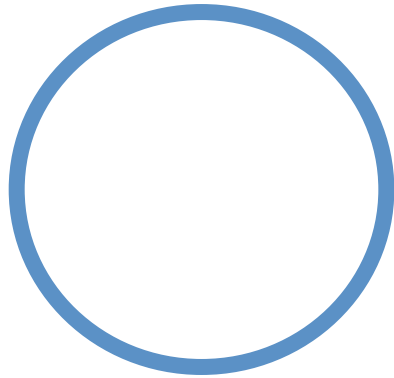
# **Finished** Prokaryotic Genome Assemblies From a **Low-cost** Combination of Short and Long Reads

- An ALLPATHS-LG recipe

Shuangye Yin

FSAF, June 2012

# Limitations of draft genome assemblies



Perfect



good but drafty

## Does it matter?

1. Mutations lost in errors
2. Gaps take out genes
3. Evolutionary hotspots missing

Manual finishing == \$\$\$

# Finished $\neq$ perfect

## Finished genomes

(Manually finished using Sanger Chemistry)

*E. coli*

*S. pneumoniae*

*R. sphaeroides*

## Reference Errors

~4

~40

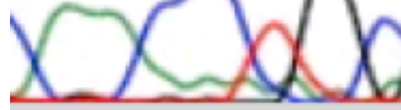
~400

Align Illumina data to the reference, find discrepancies.

For each discrepancy we examined the **original** Sanger-chemistry traces

## Deep dive into *S. pneumoniae*

### Example



miscalled as C, should be CC  
(several similar reads here)

Illumina data (our sample)  
overwhelming (CC=48, C=0)

# Affordable perfection

**Get close to perfect without breaking the bank**

## Strategy

- everything automated
- match the lab technologies to the problem



# Laboratory “recipe” / strategy

## Ingredients

50x

Illumina short pairs  
100 base reads  
from 180 bp fragments

50x

PacBio long reads  
1000 base reads  
from 2-3 kb fragments

50x

Illumina wide jumps  
100 base reads  
from 2-10 kb fragments

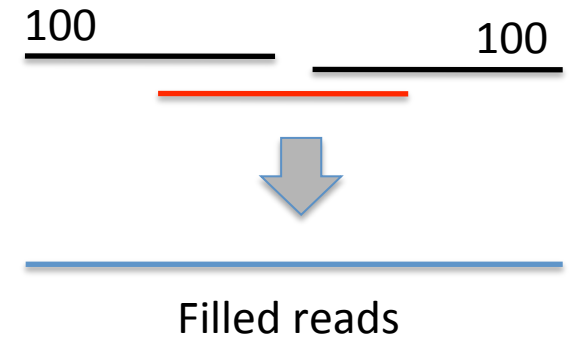
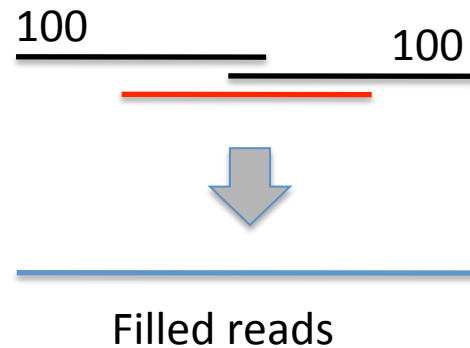
## Features

- resolve short-range repeats
- provide base accuracy
- resolve medium-range repeats
- compensate for bias
- resolve long-range repeats

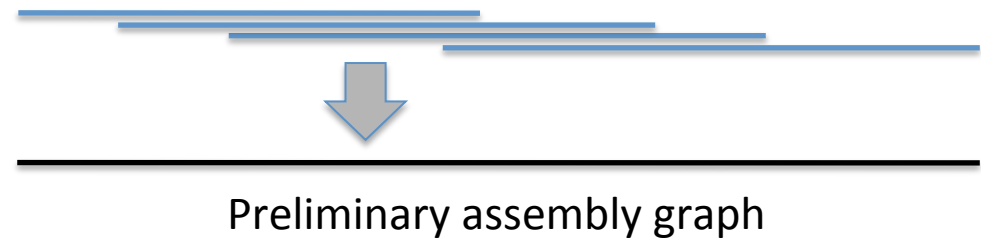
**ALGORITHM FOLLOWS THIS**

# First form initial assembly

1. **Close read pairs from 180 bp fragments**  
(3<sup>rd</sup> read – different pair)



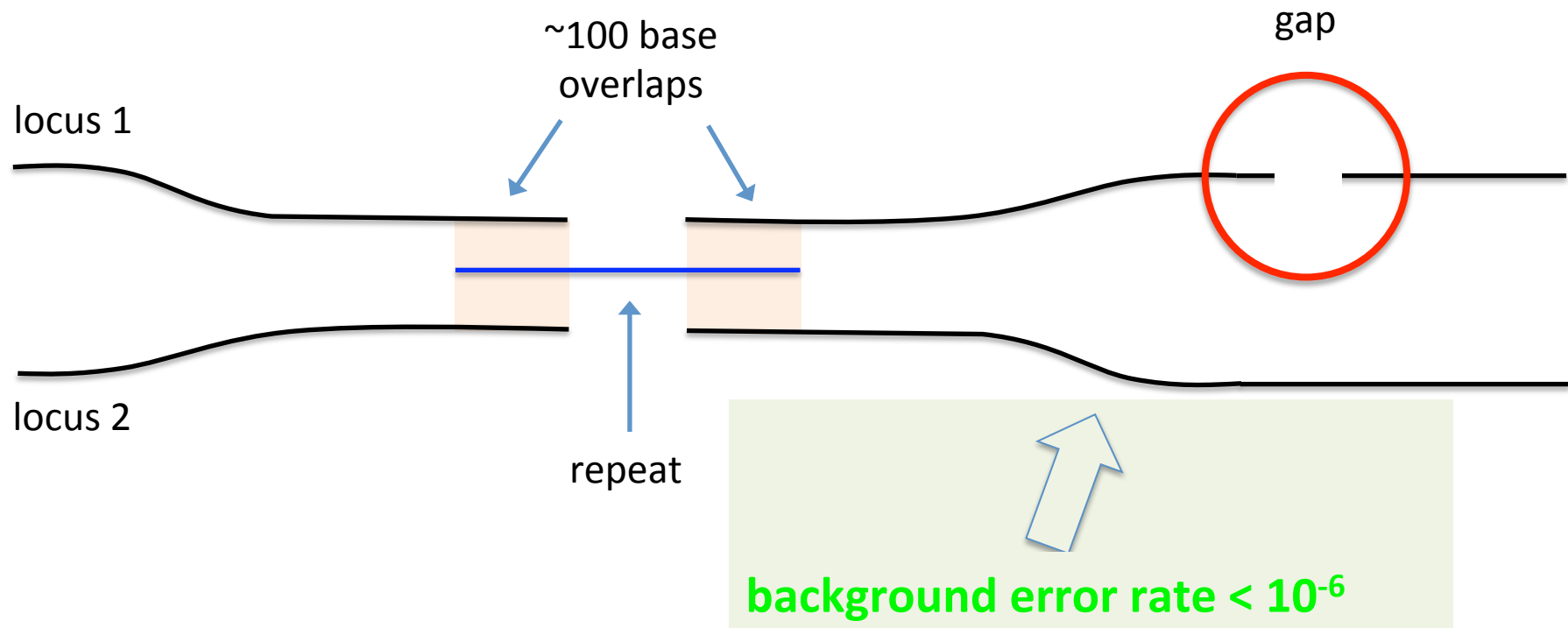
2. **Glue along ~100 base overlaps**  
(note 100  $\approx$  half of fragment size)



# Get preliminary assembly graph

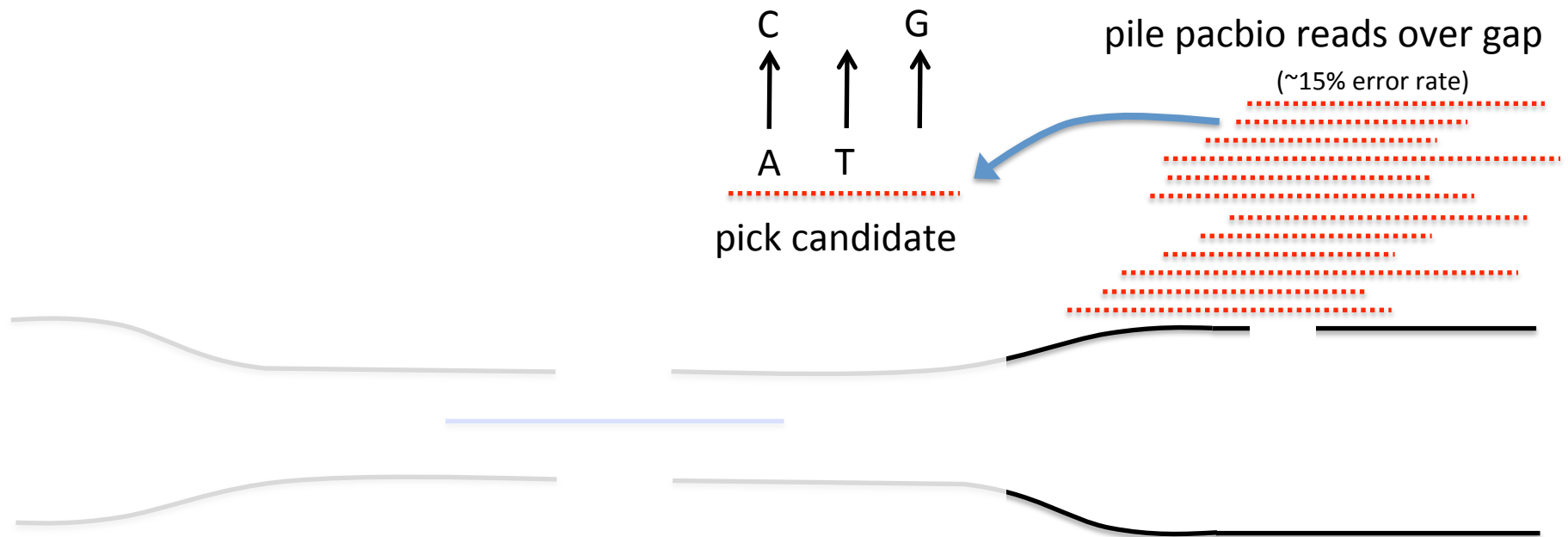
## Challenges:

- Different loci joined along repeats
- Gaps from bias



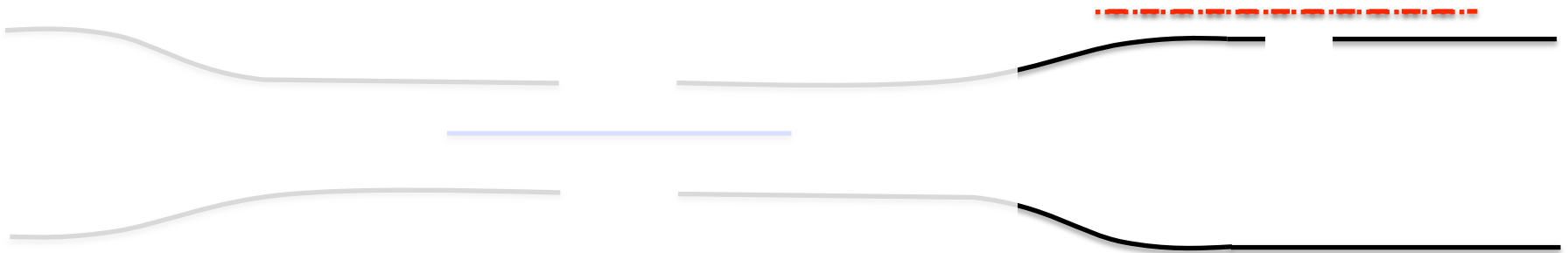


# Close graph gaps using long reads



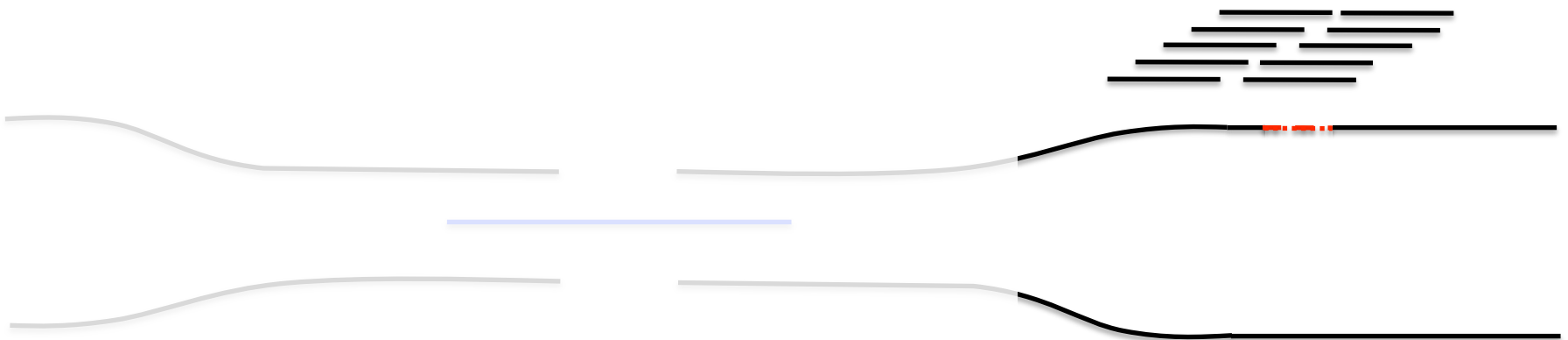
# Patches have errors

consensus patch  
has ~1% error rate



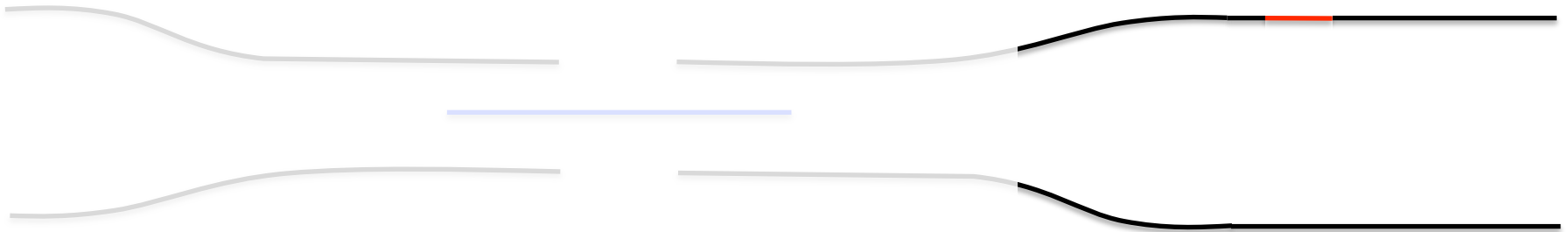
# Improve patch quality

correct patch with frag pairs

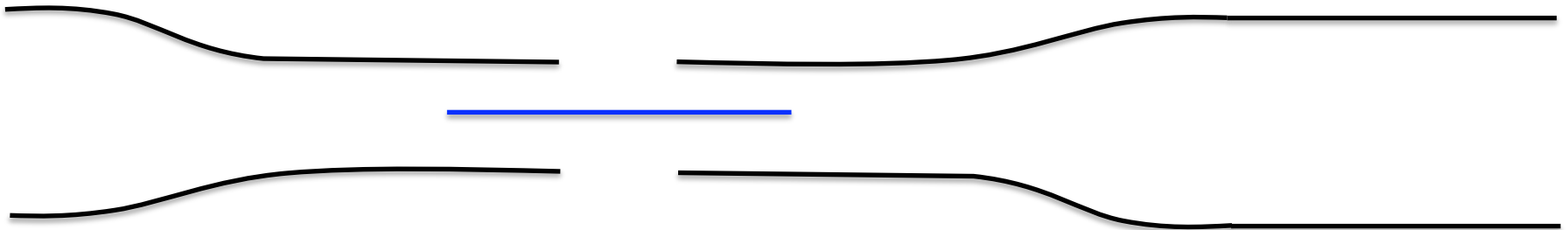


# Patches now highly accurate

nearly all patches perfect

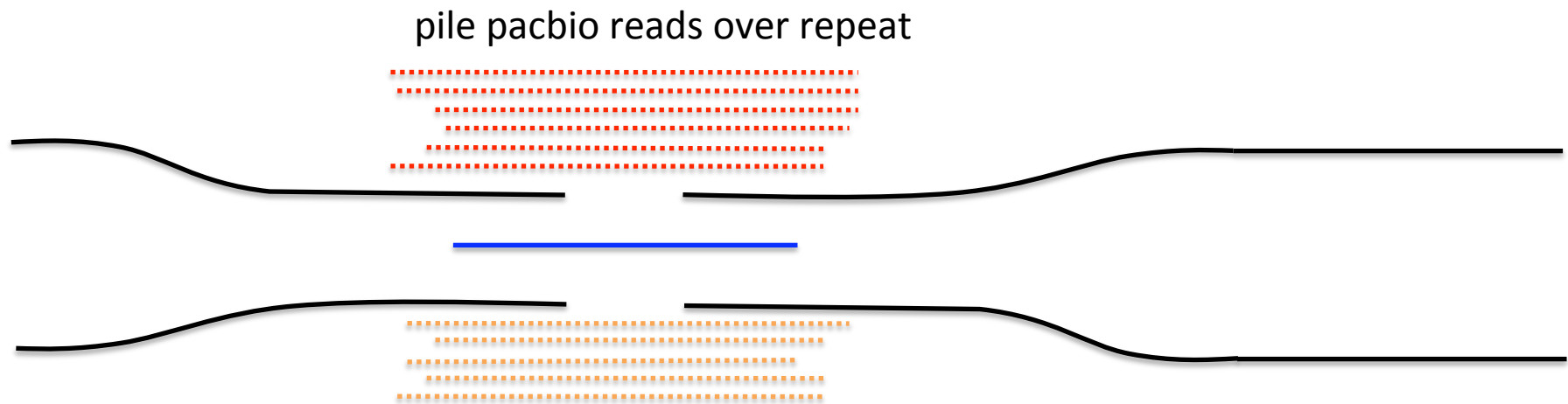


# Gaps are gone!



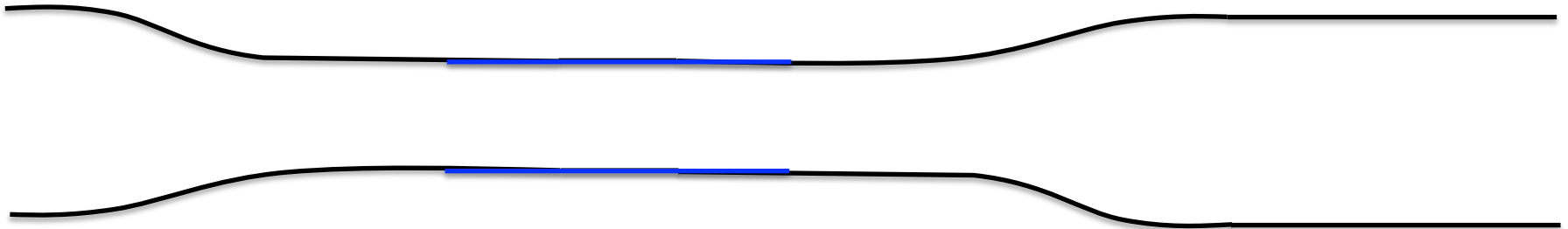
background error rate  $< 10^{-6}$

# Now disambiguate repeats



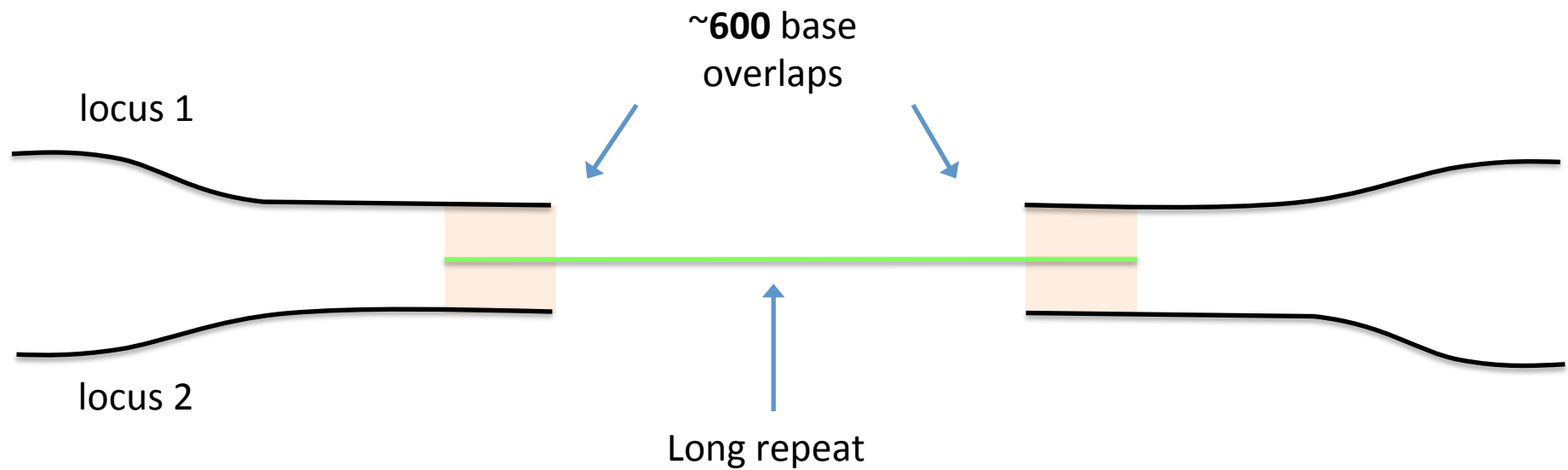
- Each pacbio read is expressed as a sequence in the graph
- Then we form the consensus of these sequences

Repeat is gone!



Same problem as before, at larger scale

ZOOM OUT!





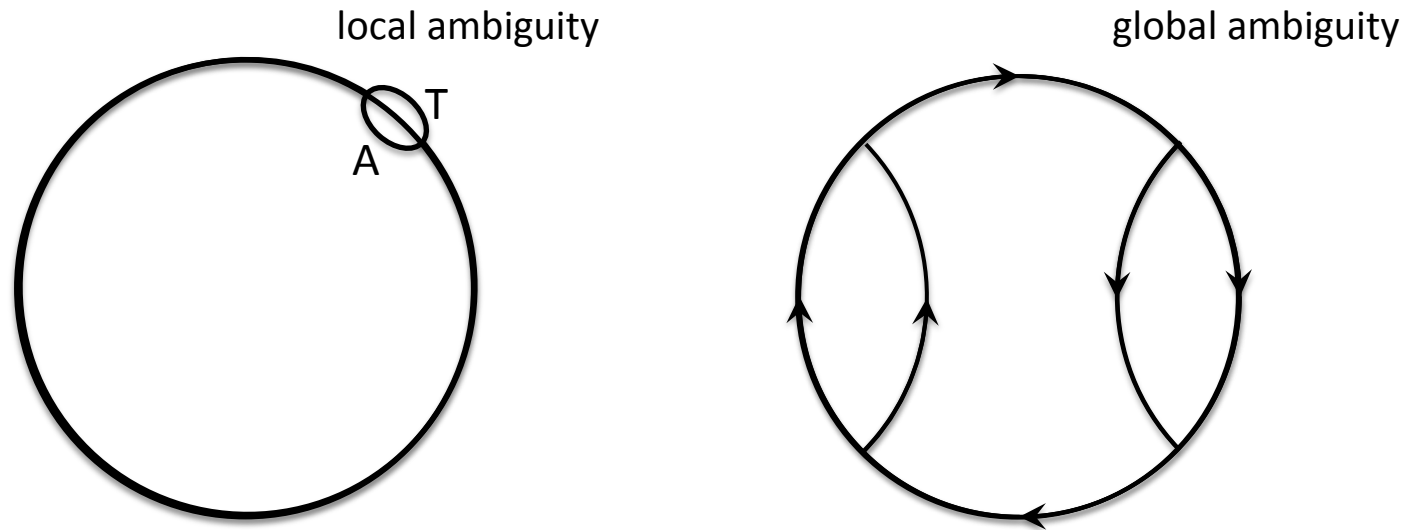
# Resolve using wide jumps



# Long repeat is gone!



# Assembly can still have ambiguities



## **FASTG: assembly format in progress**

- by Assemblathon group
- very general
- looks like FASTA

Example:

```
..CCAT[alt|A,T]GCGT..
```

# Data sets for assembly experiment

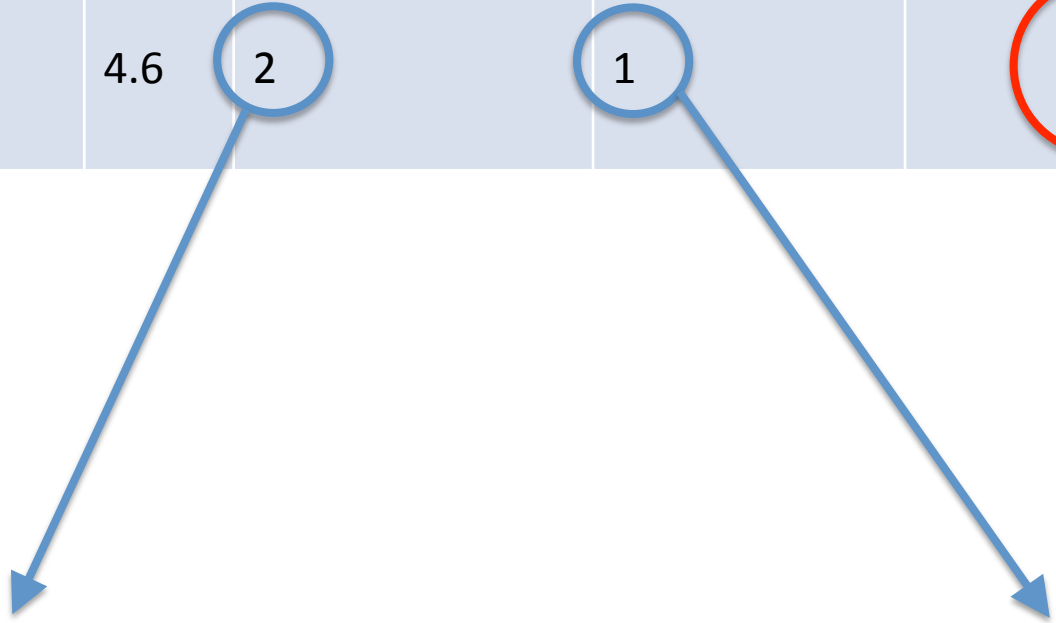
#	Species	Strain	Reference sequence
1	<i>Escherichia coli</i>	K12 MG1655	finished
2	<i>Rhodobacter sphaeroides</i>	2.4.1	finished
3	<i>Streptococcus pneumoniae</i>	Tigr4	finished
4	<i>Bacteroides eggerthii</i>	1_2_48FAA	
5	<i>Bacteroides fragilis</i>	CL05T00C42	
6	<i>Bacteroides thetaiotaomicron</i>	CL09T03C10	
7	<i>Bifidobacterium bifidum</i>	NCIMB 41171	
8	<i>Coprobacillus</i> sp.	D6	
9	<i>Enterococcus casseliflavus</i>	EC20	
10	<i>Eubacterium</i> sp.	3_1_31	
11	<i>Fusobacterium nucleatum</i>	OT 420	
12	<i>Fusobacterium nucleatum</i>	7_1	
13	<i>Klebsiella oxytoca</i>	10-5248	
14	<i>Neisseria gonorrhoeae</i>	FA19	
15	<i>Neisseria gonorrhoeae</i>	MS11	
16	<i>Scardovia wiggisiae</i>	F0424	

GC content from 27% to 69%.

- data generated by same automated recipe.
- assemblies run with same parameters

# ALLPATHS-LG assemblies of finished genomes

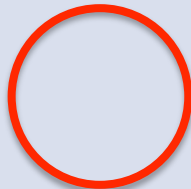

species	size (Mb)	errors	ambiguities	assembly (to scale)
<i>Escherichia coli</i>	4.6	2	1	



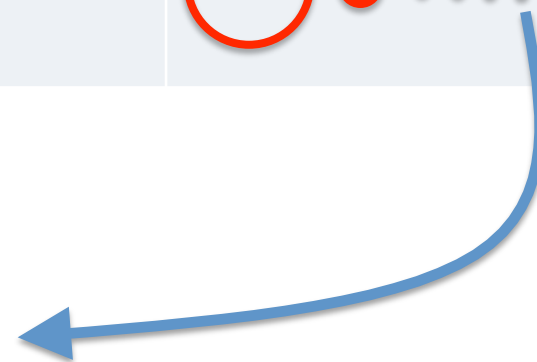
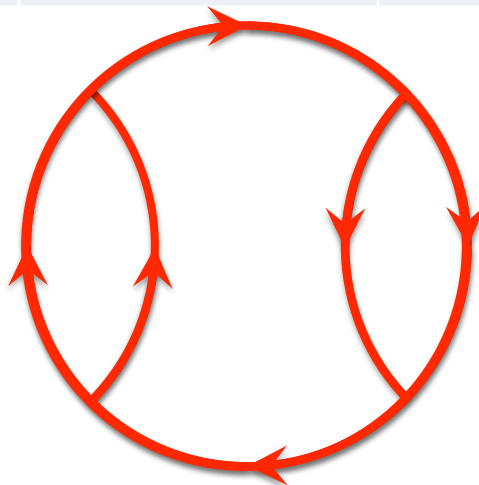
two substitutions  
separated by three  
bases

substitution

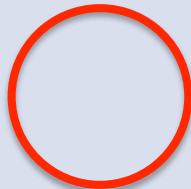


# ALLPATHS-LG assemblies of finished genomes

species	size (Mb)	errors	ambiguities	assembly (to scale)
<i>Escherichia coli</i>	4.6	2	1	
<i>Rhodobacter spheroides</i>	4.6	4	9	

two plasmids intertwined  
along 15 kb repeats







# ALLPATHS-LG assemblies of finished genomes

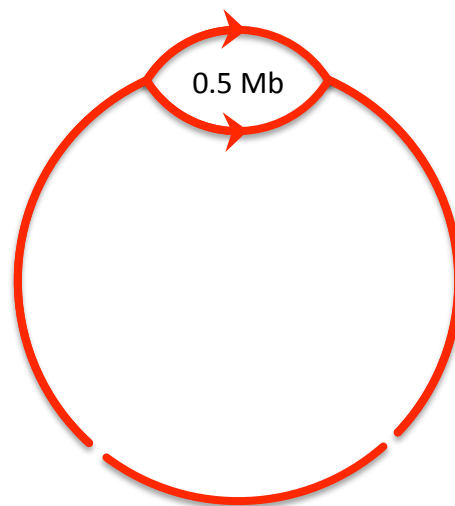
species	size (Mb)	errors	ambiguities	assembly (to scale)
<i>Escherichia coli</i>	4.6	2	1	
<i>Rhodobacter spheroides</i>	4.6	4	9	
<i>Streptococcus pneumoniae</i>	2.2	0	6	

Reference errors

The stats are better than we gave for finished sequences!

# Other nearly complete assemblies

species	size (Mb)	ambiguities	assembly (to scale)
<i>Bifidobacterium bifidum</i>	2.2	4	
<i>Scardoviawiggisiae</i>	1.5	2	
<i>Enterococcus casseliflavus</i>	3.4	0	
<i>Eubacterium</i> sp.	3.1	15	



bubble flanked by 13 kb inverted repeat

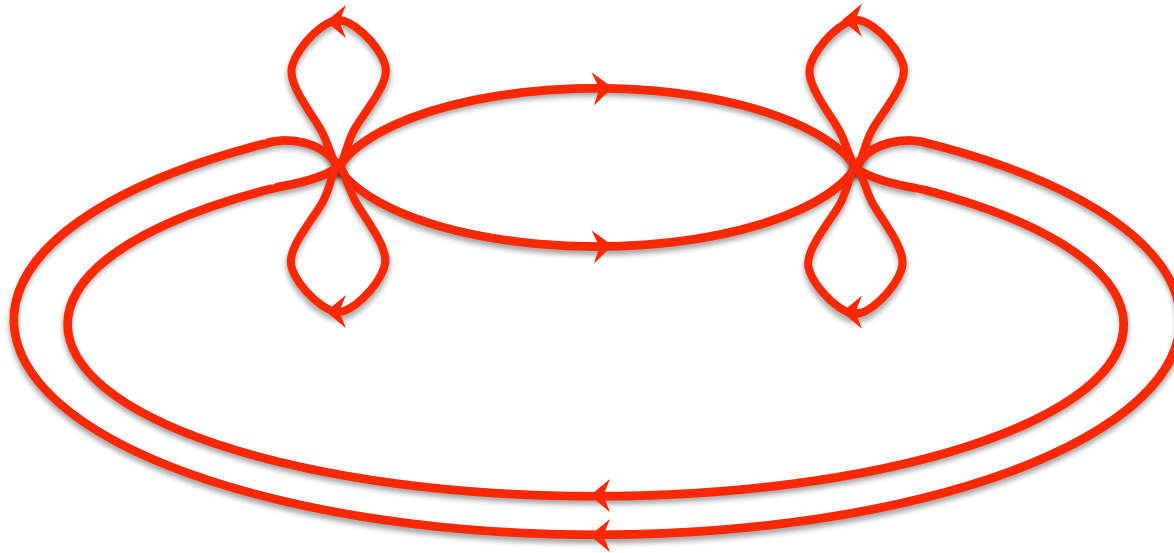


# Messier assemblies

Other assemblies: less well resolved

Some have several gaps and some are tangled

***Bacterioides thetaiotaomicron*** (example)



What's happening: repeat occurs eight times, half in reverse orientation. Long enough that power may be lacking to pull it apart.

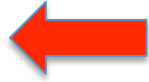
# Why are some assemblies messier?

Likely causes:

- repeat sizes vary
- jump libraries vary

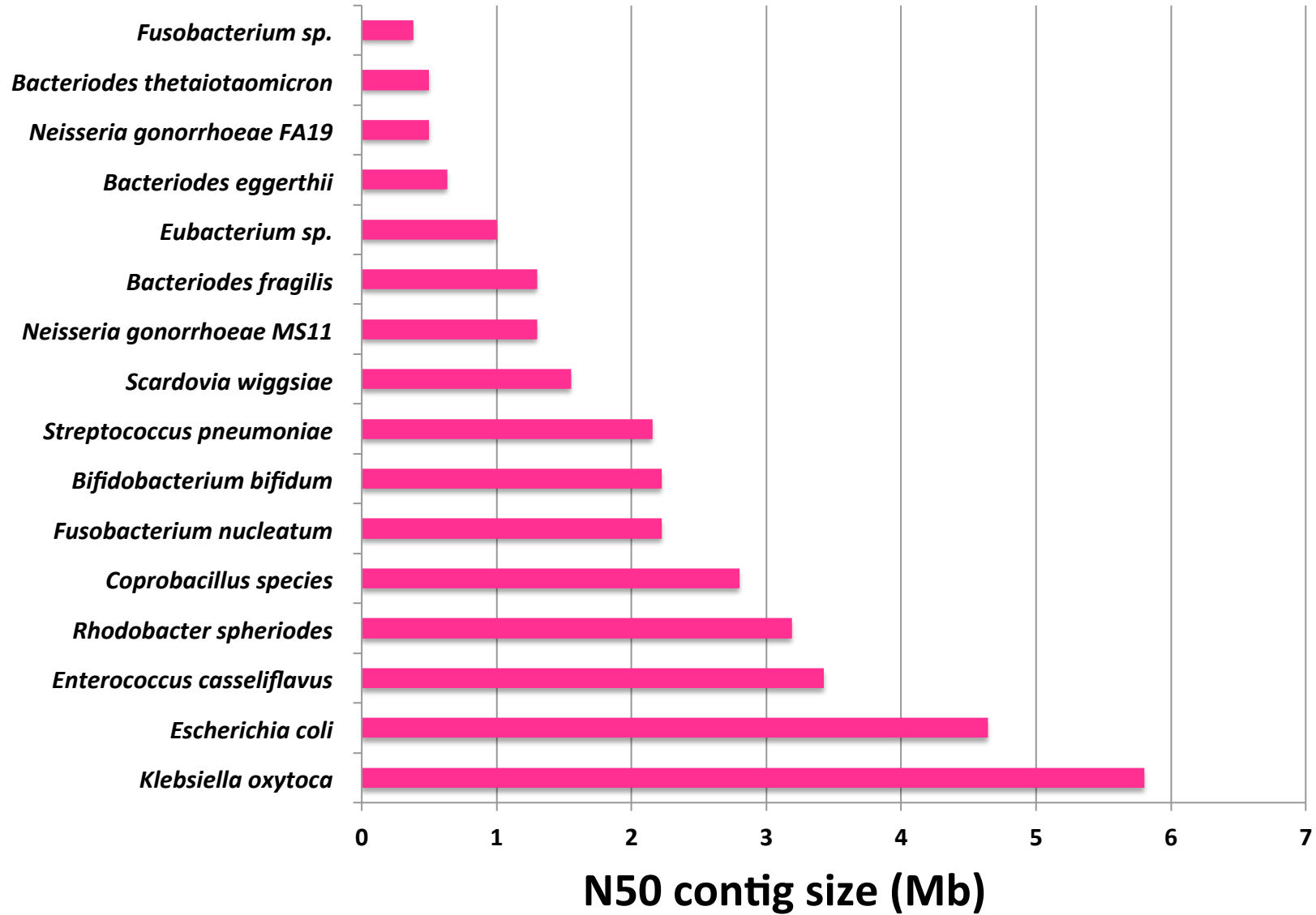
**Average number of jumps covering a window of given size**

sample	1k	2k	3k	4k	5k	6k
1	229	141	75	37	17	8
2	262	141	65	27	11	4
3	272	159	79	36	16	7
4	153	75	32	12	4	1.3
5	287	148	60	21	6	1.7
6	449	191	58	14	2	0.5
7	256	158	84	40	17	8
8	396	198	81	28	8	2
9	278	131	50	17	5	1.3
10	50	28	13	6	3	1.1
11	304	141	51	15	4	0.8
12	243	116	46	16	5	1.4
13	573	285	114	39	12	3
14	436	228	99	37	13	4
15	424	258	139	69	33	15
16	435	185	56	12	2	0.2

 75-fold variability between jump power

Tried manually increase the cover by 2.5 fold => Much better assembly.

# Our contigs are really big





## **Near perfect assembly of bacterial genomes**

- High quality genome, cost far lower
- Methods (lab + ALLPATHS-LG) available
- We're here to help

# Acknowledgments

---

## **Broad Institute**

### Computational

David Jaffe  
Iain MacCallum  
Dariusz Przybylski  
Filipe Ribeiro  
Michael Ross  
Ted Sharpe  
Sante Gnerre  
Terry Shea  
Bruce Walker  
Sarah Young

### 180 bp

Brian Hurhula  
Chris Friedrich  
Cole Walsh  
Danielle Perrin  
Sheila Fisher

### Jumps

Marc Chevrette  
Purnima Kompella  
Riza Daza  
Scott Steelman

### PacBio

James Meldrim  
Brian Sogoloff  
Patrick Cahill  
Todd Sparrow  
Lynne Aftuck

Carsten Russ  
Nick Patterson  
Rob Nicol  
Chad Nusbaum

## **Funding**

**NHGRI**

**NIAID**

Many colleagues who contributed DNA samples